On the Characterization of Protein Native State Ensembles

Amarda Shehu,* Lydia E. Kavraki,*†§ and Cecilia Clementi‡§

Departments of *Computer Science, †Bioengineering, and ‡Chemistry, Rice University, Houston, Texas; and §Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas

ABSTRACT Describing and understanding the biological function of a protein requires a detailed structural and thermodynamic description of the protein's native state ensemble. Obtaining such a description often involves characterizing equilibrium fluctuations that occur beyond the nanosecond timescale. Capturing such fluctuations remains nontrivial even for very long molecular dynamics and Monte Carlo simulations. We propose a novel multiscale computational method to exhaustively characterize, in atomistic detail, the protein conformations constituting the native state with no inherent timescale limitations. Applications of this method to proteins of various folds and sizes show that thermodynamic observables measured as averages over the native state ensembles obtained by the method agree remarkably well with nuclear magnetic resonance data that span multiple timescales. By characterizing equilibrium fluctuations at atomistic detail over a broad range of timescales, from picoseconds to milliseconds, our method offers to complement current simulation techniques and wet-lab experiments and can impact our understanding and description of the relationship between protein flexibility and function.

INTRODUCTION

It is well established that, while an experimentally determined structure may reveal a protein's functional regions, structural fluctuations under native conditions can modulate function (1–3). Experiments, simulations, and theory indicate that a detailed description of function (encompassing enzymatic reactions, electron transfer, protein ligand binding, and protein/protein interactions) requires the characterization of a protein's native state as an ensemble of conformations (4–7). Such a characterization involves describing in detail the structural and thermodynamic properties over all conformations of the native state ensemble.

Obtaining this description has proven challenging. While nuclear magnetic resonance (NMR) spectroscopy describes picosecond-millisecond timescale dynamics through relaxation phenomena (1,8,9), the characterization of all the conformations constituting the native state at atomistic detail remains an active area of research (10). Molecular dynamics (MD) and Monte Carlo (MC) methods, especially when combined with enhanced sampling techniques and massive parallelization (11–14) or when conducted in low-dimensional configuration spaces (15–17), are powerful complements to characterize the native state ensemble (18). However, the computational demand of these methods makes it challenging to explore longer timescales (19–21). Efforts to explore native state ensembles with no timescale limitations have recently focused either on obtaining native thermodynamic propensities of amino acids (22) or on generating conformations of the native state ensemble by guiding MD or MC with explicit information from NMR measurements (10,23,24).

In this context, we have recently developed the Protein Ensemble Method (PEM) (25) to exhaustively characterize the native state ensemble of a protein at atomistic detail with no inherent timescale limitations. PEM obtains all-atom conformations of the native state in a multiscale fashion combining geometric and energetic considerations. On the generated conformations, PEM measures thermodynamic averages in a statistical mechanics framework and so allows a direct quantitative comparison with wet-lab experimental measurements. We have shown that PEM is intrinsically parallel, efficient in generating large ensembles, and able to characterize equilibrium fluctuations of both loop segments and polypeptide chains (25,26).

In this work, we show the generality of PEM by using the method to characterize native state ensembles of proteins of different sizes and folds. We present the PEM-obtained native state ensembles of eglin c, the SH3 domain of Fyn tyrosine kinase (FynSH3), the 10th type III domain of fibronectin (FNfn10), and the *Peptostreptococcus magnus* albumin-binding second GA module of PAB (ALB8-GA). These proteins are 70, 58, 90, and 53 aa long, of $\alpha + \beta$, mainly β, all β, and all α-folds, respectively. We show that for all these proteins the PEM-obtained native fluctuations agree remarkably well with NMR data such as order parameter and three-bond scalar coupling data. In addition, for ALB8-GA, where side-chain NMR data are presently not available, we present our prediction on equilibrium side-chain fluctuations.

MATERIALS AND METHODS

We first briefly review the main components of PEM. A more detailed discussion of the method can be found in Shehu et al. (26).

Generation of native state ensembles

PEM employs the following multi-scaling approach to generate the native state ensemble of a protein:

Submitted July 31, 2006, and accepted for publication November 13, 2006.

Address reprint requests to Lydia E. Kavraki, Tel.: 713-348-5737; E-mail: kavraki@rice.edu; or address reprint requests to Cecilia Clementi, Tel.: 713-348-3485; E-mail: cecilia@rice.edu.

0006-3495/07/03/1503/09 $2.00

doi: 10.1529/biophysj.106.094409

1. Starting from the topology of an initial native structure (used as a reference), the method first divides the polypeptide chain into consecutive long segments of significant overlap.
2. For each segment, an extensive ensemble of relevant backbone configurations is obtained through a geometric exploration of conformational space that combines uniform sampling of the backbone dihedral degrees of freedom of the segment with an efficient inverse kinematics procedure known as cyclic coordinate descent (27).
3. Optimal side-chain configurations are then added onto each backbone configuration, and a short energy minimization of each of the resulting all-atom conformations is finally performed. A generated conformation is deemed low-energy and added to the native state ensemble if its energy is no higher than 20 kcal/mol from the energy of the initial structure employed.

Equilibration of solution structures

For the proteins presented here, an initial native structure is obtained by equilibrating an NMR solution structure. NMR ensembles of solution structures of eglin c (28), FynSH3 (29), FNfn10 (30), and ALB8-GA (31) are available in the PDB (32) under codes 1egl, 1nyg, 1ttf, and 1gab. The solution structure that is reported as the best, representative, or the average of the NMR ensemble for each protein is subjected to a short energy minimization. The average structures of the NMR ensembles of FynSH3, FNfn10, and ALB8-GA are reported under PDB codes 1nyf, 1ttg, and 1prb. When a best, representative, or average structure is not reported in the PDB, which is the case for eglin c, the first structure of the NMR ensemble is chosen to be subjected to an energy minimization procedure.

The energy of a structure is measured through the CHARMM all-atom force field (33). The energy minimization procedure involves a conjugate gradient descent in the energy landscape. The minimization of a structure is considered converged if during the last 300 steps of the conjugate gradient descent the improvement in energy is <2.0 kcal/mol. Equilibrated structures of eglin c, FynSH3, FNfn10, and ALB8-GA differ from their corresponding solution structures by all-atom RMSDs of 1.8, 1.7, 2.0, and 2.5 Å, respectively (the effect of the equilibration of PDB-obtained structures on the native state ensembles generated by PEM is discussed in full in Shehu et al. (26)).

PEM divides the polypeptide chain of each of these proteins into segments of 30 aa long with an overlap with each other of 25 aa. The values for the segment length and overlap are chosen by a general and automated procedure. Optimal segment length and overlap result in consistent amino acid fluctuations as measured over the ensembles generated for overlapping segments enclosing each amino acid (see (26) for details and for values to all parameters used by PEM).

Measurement of thermodynamic averages

PEM measures thermodynamic averages over the segment ensembles in a statistical mechanics framework. Each PEM-generated conformation C with energy $E(C)$ is weighted by its Boltzmann probability $P(C) = P_{\text{ref}}\, e^{-(\text{E(C)}-\text{E}_{\text{ref}})/\text{RT}_0}$, where E_{ref} is the energy of the equilibrated solution structure (taken as reference), R is the gas constant, and T_0 is room temperature of 300 K. The constant P_{ref} is the probability of the reference structure and can be set to 1 without loss of generality. Let $X_{\text{i}}(C)$ indicate the value of an observable X, at position i, measured on conformation C; the thermodynamic average of this quantity over the generated ensemble is measured as $\langle X_{\text{i}}\rangle = \frac{1}{\text{Q}}\sum_{\text{C}} e^{-(\text{E(C)}-\text{E}_{\text{ref}})/\text{RT}_0}\, X_{\text{i}}(C)$, where Q refers to the partition function. Averages measured over ensembles of neighboring segments are then combined to obtain structural and thermodynamic observables of the native state. Since a conformation C with energy $E(C)$ higher than 20 kcal/mol from the reference energy E_{ref} has an associated relative Boltzmann probability $\frac{\text{P(C)}}{\text{P}_{\text{ref}}} \lesssim 10^{-15}$, its contribution to ensemble averages $\langle X_{\text{i}}\rangle$ is practically negligible. Therefore, only conformations whose energies are no higher than a cutoff of 20 kcal/mol from the reference energy E_{ref} are considered in the ensembles.

The thermodynamic observables calculated over the PEM-obtained ensembles consist of amide and methyl order parameter (S^2) data that measure the reorientational averaging of amide and methyl bonds, respectively, and three-bond scalar coupling (^3J) data that measure side-chain rotamer averaging. These average values can be directly compared to the corresponding values measured in NMR experiments and quantify native fluctuations of a protein at varying timescales. While amide S^2 data measure picosecond-nanosecond timescale fluctuations, methyl S^2 and ^3J data can span up to millisecond timescales (1,8,9).

S^2 data for a bond are measured by averaging over the distribution of vectors assumed by the bond in a generated ensemble (23). The calculation of S^2 data is based on the Lipari-Szabo model-free formalism (34) that does not assume a particular model of internal motions. The model-free formalism allows for a direct comparison of calculated S^2 values with experimental order parameters under the assumption that motions of the methyl symmetry axis and of the protons about this axis are decoupled (35). A thorough discussion on the model-free formalism can be found in the literature (34,35). Based on the Lipari-Szabo model-free formalism (34), the order parameter $S^2_{\text{i, j}}$ for a bond between atoms i and j is calculated through the formula

$$S^2_{\text{i,j}} = \frac{3}{2}\Big(\langle \hat{x}^2_{\text{i,j}}\rangle + \langle \hat{y}^2_{\text{i,j}}\rangle + \langle \hat{z}^2_{\text{i,j}}\rangle + 2\langle \hat{x}_{\text{i,j}}\hat{y}_{\text{i,j}}\rangle^2 + 2\langle \hat{x}_{\text{i,j}}\hat{z}_{\text{i,j}}\rangle^2 + 2\langle \hat{y}_{\text{i,j}}\hat{z}_{\text{i,j}}\rangle^2 - \frac{1}{2}\Big),$$

where $\hat{x}, \hat{y}, \hat{z}$ denote the components of the unit vector along the bond. Since bond lengths remain essentially unchanged from their equilibrium values during PEM's execution, the above formula can be simplified as in Best and Vendruscolo (23) to

$$S^2_{\text{i,j}} = \frac{3}{2(r^{\text{min}}_{\text{i,j}})^4}\Big(\langle x^2_{\text{i,j}}\rangle + \langle y^2_{\text{i,j}}\rangle + \langle z^2_{\text{i,j}}\rangle + 2\langle x_{\text{i,j}}y_{\text{i,j}}\rangle^2 + 2\langle x_{\text{i,j}}z_{\text{i,j}}\rangle^2 + 2\langle y_{\text{i,j}}z_{\text{i,j}}\rangle^2 - \frac{1}{2}\Big),$$

where $r^{\text{min}}_{\text{i,j}}$ refers to the equilibrium length of the bond connecting atoms i and j. The ensemble-averaged S^2 for a particular bond is thus obtained by Boltzmann-averaging over the distribution of x, y, z components of vectors assumed by the bond. $S^2 = 1$ indicates no heterogeneity in the distribution of these vectors, whereas $S^2 = 0$ is indicative of a uniform distribution. Similarly, ^3J data are measured over the distribution of assumed rotamers (36). The calculation of these quantities and their comparison with NMR data allows us to quantitatively assess the agreement between the PEM-generated and the actual native state ensembles.

Additional measurements presented in this work consist of probabilities of contacts and hydrogen bonds, which are similarly Boltzmann-weighted. Two amino acids are considered in contact with one another if the Euclidean distance between two of their atoms is no more than 4.5 Å. A hydrogen bond is considered formed if the OH distance is <2.4 Å and the maximum NHO angle for the hydrogen bond alignment is 2.44 rad.

The computational uncertainty associated with the thermodynamic observables calculated over the PEM-generated ensembles is obtained by measuring differences in the observables when alternative implementation decisions are made in PEM. Therefore, the error bars associated with the PEM-calculated thermodynamic observables measure the inherent error, hence the robustness, of PEM (see (26) for a list of all implementation decisions).

The Pearson correlation R^2 and reduced χ^2 are used to quantify the agreement between calculated and experimental thermodynamic averages. They are measured as defined in Bevington and Robinson (37).

Computational cost

For each of the proteins in this study, ~13,000 conformations with energy within 20 kcal/mol from the reference structure are generated for each 30 aa segment. Of these, ~5000 conformations per segment have energies no higher than 5 kcal/mol from the energy of the equilibrated solution structure

used as reference. All results presented here were obtained on the Rice University Terascale cluster of 900 MHz Intel Itanium2 processors (Intel, Santa Clara, CA) and on the Rice University ADA cluster of 2.2 GHz AMD Opteron processors. The calculations for each protein required <100 CPU hours.

RESULTS

Figs. 1 *a*, 2 *a*, 4 *a*, and 6 *a*, show the obtained conformational ensembles for eglin c, FynSH3, FNfn10, and ALB8-GA, respectively. Figs. 1 *b*, 2 *b*, 4 *b* and *c*, and 6 *b*, show that correlations between the S^2 and ^3J data calculated over the ensembles obtained for eglin c, FynSH3, and FNfn10 (S^2_{calc} and $^3J_{calc}$) and the NMR S^2 and ^3J data (S^2_{exp} and $^3J_{exp}$) are >92%. This result is particularly significant when considering the low correlations, 37–50%, between the S^2_{exp}, $^3J_{exp}$ data and the corresponding quantities measured over the NMR ensembles (28–31) available for these proteins. Results for each protein are discussed in the following.

Analysis of PEM-generated native state ensemble of eglin C

Fig. 1 *a* shows the native state ensemble obtained by PEM for eglin c. Fig. 1 *a* clearly shows the heterogeneity of this ensemble. The largest equilibrium fluctuations obtained for this protein are located in the Thr[1]-Gly[15] N-terminus, which is practically disordered. Interestingly, the protease-binding loop, encompassing amino acids Ser[41]-Arg[48], is also very mobile. Of all the amino acids of the loop, Val[43]-Leu[47] are the most mobile. The mobility of the entire loop is also reflected in the low average of 0.7 of the amide S^2_{calc} data corresponding to the amide bonds of the loop's amino acids.

The entire amide and methyl S^2_{calc} data computed over the ensemble obtained for eglin c are shown in Fig. 1 *b*. Fig. 1 *b* shows that S^2_{calc} agree with S^2_{exp} data (38) with a Pearson correlation of 95% and reduced χ^2 of 0.98. Methyl S^2_{calc} data measured over the generated native state ensemble of eglin c are on average as low as 0.49. This is mostly due to the disordered Thr[1]-Gly[15] N-terminus.

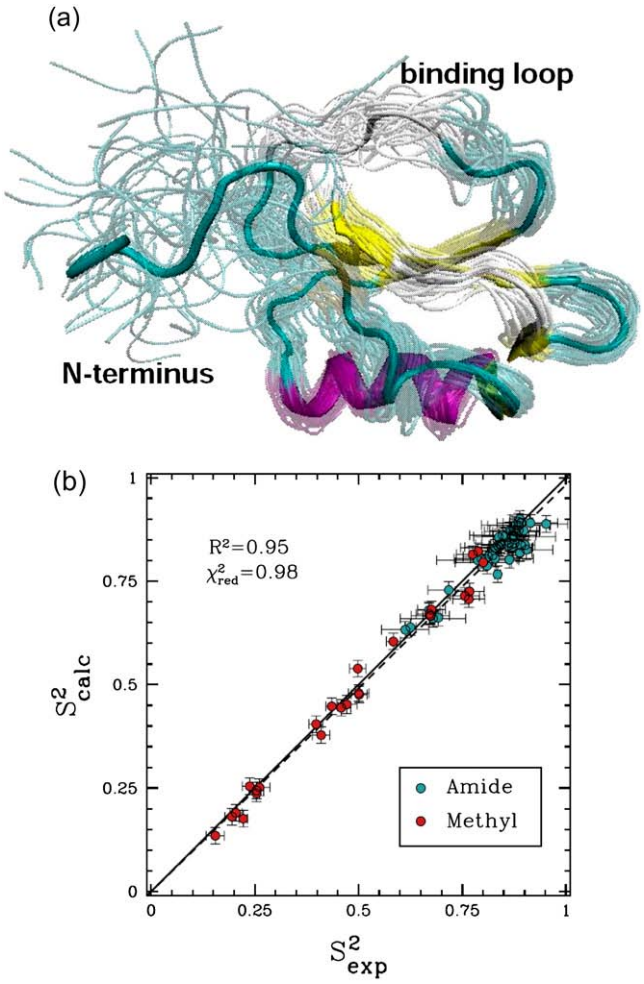

FIGURE 1 (*a*) Eglin c conformations with energy no higher than 5 kcal/mol from the equilibrated solution structure, shown as opaque, are drawn in transparent representation. (*b*) Calculated amide and methyl S^2 data (S^2_{calc} on the *y* axis) are compared to NMR S^2 data (S^2_{exp} on the *x* axis). The dashed line indicates the linear least squares regression fit on the data sets. The solid line is the identity line.

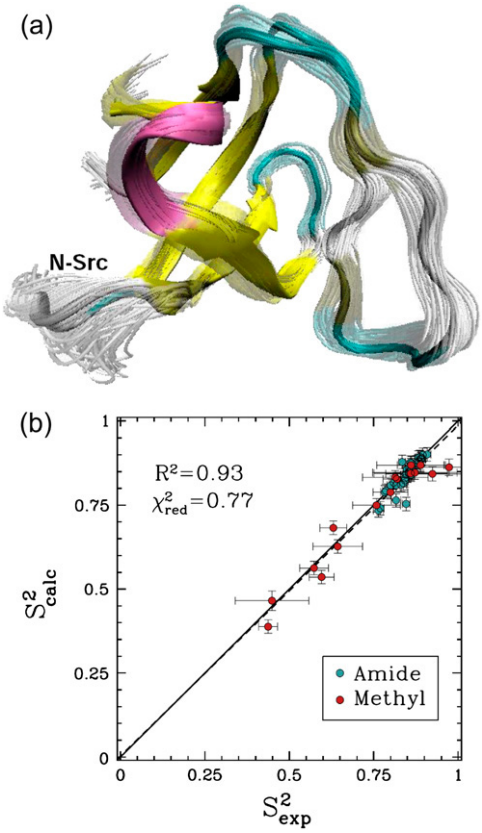

FIGURE 2 (*a*) Fyn SH3 conformations with energy no higher than 5 kcal/mol from the equilibrated solution structure, shown as opaque, are drawn in transparent representation. (*b*) Calculated amide and methyl S^2 data (S^2_{calc} on the *y* axis) are compared to NMR S^2 data (S^2_{exp} on the *x* axis). The dashed line indicates the linear least squares regression fit on the data sets. The solid line is the identity line.

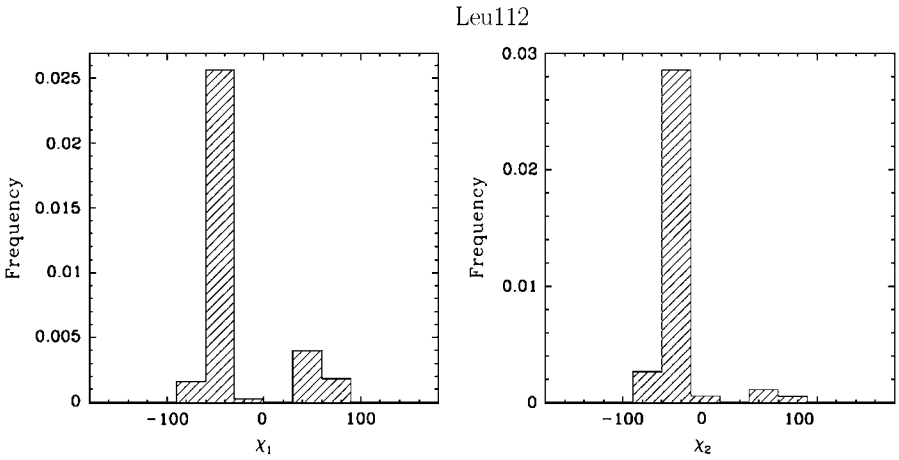

FIGURE 3 Distributions of χ_1 and χ_2 angles (χ_1 and χ_2 correspond to the dihedral angles associated with the $C_\gamma - C_{\delta_1}$ and the $C_\gamma - C_{\delta_2}$ bonds, respectively) for Leu^{112} in FynSH3 reveal that Leu^{112} prefers more than one rotameric state.

Analysis of PEM-generated native state ensemble of Fyn SH3

The obtained native state ensemble of FynSH3 is shown in Fig. 2 *a*. In contrast to the ensemble obtained for eglin c, Fig. 2 *a* shows that the obtained equilibrium fluctuations for FynSH3 are prevalently small-scale. The largest fluctuations are located in the N-Src loop, which encompasses amino acids Asn^{113}-Trp^{119}. Interestingly, the N-Src loop discriminates between class I and class II ligands binding to FynSH3 (29). Of all this loop's amino acids, its central amino acid, Glu^{116} is the most mobile.

The obtained equilibrium fluctuations of FynSH3 are validated by comparing S^2_{calc} data to the corresponding S^2_{exp} NMR data (39). Fig. 2 *b* shows that S^2_{calc} and S^2_{exp} data (39) for FynSH3 agree with a Pearson correlation of 93% and reduced χ^2 of 0.77. The small-scale fluctuations qualitatively shown in Fig. 2 *a* are reflected in the S^2_{calc} data: amide and

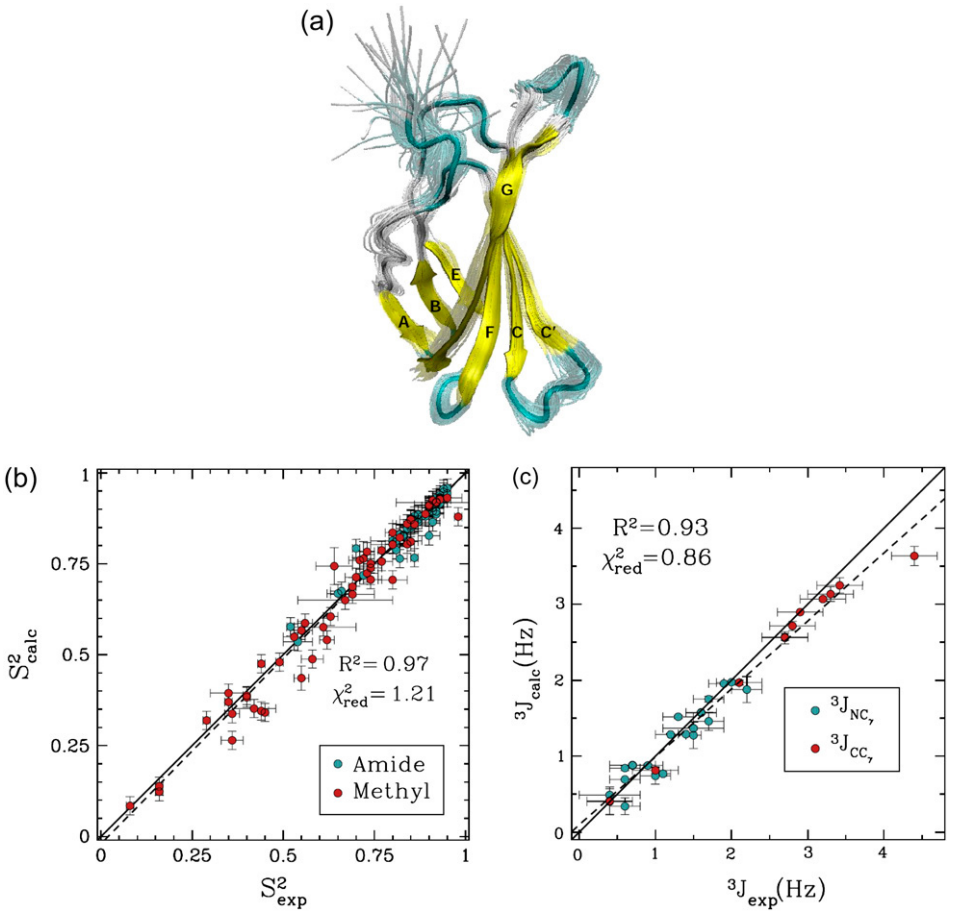

FIGURE 4 (*a*) FNfn10 conformations with energy no higher than 5 kcal/mol from the equilibrated solution structure, shown as opaque, are drawn in transparent representation. (*b*) Calculated amide and methyl S^2 data (S^2_{calc} on the *y* axis) are compared to NMR S^2 data (S^2_{exp} on the *x* axis). (*c*) Calculated $^3J_{NC_\gamma}$ and $^3J_{CC_\gamma}$ ($^3J_{calc}$ on the *y* axis) are compared to NMR ^3J data ($^3J_{exp}$ on the *x* axis). (*b* and *c*) The dashed black line indicates the linear least-squares regression fit on the data sets. The continuous line is the identity line.

methyl S^2_{calc} data have high averages of 0.84 and 0.72. This result agrees with experimental findings that large amplitude microsecond-millisecond motions are unlikely in the FynSH3 native state (39).

An interesting instance is represented by amino acid Leu^{112}, located at the border between a β-sheet and the beginning of the N-Src loop. The methyl S^2_{calc} values associated with the χ_1 and χ_2 angles of Leu^{112} are the lowest in the whole protein, even though the backbone fluctuations at this position are limited. Fig. 3 shows the distribution of the side-chain χ_1 and χ_2 angles in Leu^{112} and reveals that the low methyl S^2_{calc} data result from averaging over multiple rotameric states populated by the side chain of Leu^{112} in the ensemble.

Analysis of PEM-generated native state ensemble of FNfn10

The native state ensemble obtained for FNfn10 is shown in Fig. 4 *a*. The N-terminal amino acids appear disordered, while the seven β-strands of FNfn10, A, B, C, C′, E, F, and G, are well defined and practically rigid. The surface loops connecting the β-sheets (AB, BC, CC′, C′E, EF, and FG), however, are shown to be mobile. The PEM-obtained mobility for these loops agrees with the hypothesis that motions of these loops play a role in the induced-fit recognition of FNfn10 by multiple receptors (40). In particular, the most mobile amino acids, Val^{27}, Ser^{43}, and Arg^{78}, are located in the BC, CC′, and FG loops. Interestingly, the FG loop, which includes the RGD cell-adhesion motif, encompassing amino acids Arg^{78}-Asp^{80} (40), is the most flexible of all the surface loops in FNfn10.

Fig. 4, *b* and *c*, show that S^2_{calc} and $^3J_{calc}$ for FNfn10 agree with S^2_{exp} and $^3J_{exp}$ data (41) with Pearson correlations of 97% and 93%, and reduced χ^2s of 1.21 and 0.86, respectively. Amide S^2 data with a high average of 0.86 indicate small-scale fluctuations and a practically rigid hydrophobic core. This result agrees with the findings reported in Carr et al. (40), where microsecond-millisecond motions in FNfn10 are not observed.

While most side chains have a single staggered rotamer, Val^4, Val^{11}, and Val^{50} have unusually low ^3J values, indicative of rotamer averaging. Distributions of the side-chain γ_1 and γ_2 angles in these amino acids are measured over the obtained native state ensemble of FNfn10 and shown in Fig. 5. Fig. 5 confirms that Val^4, Val^{11}, and Val^{50}, while preferring one rotamer, are found on average in 4–5 rotamers.

Analysis of PEM-generated native state ensemble of ALB8-GA

The native state ensemble obtained by PEM for ALB8-GA is shown in Fig. 6 *a*. Fig. 6 *b* shows the amide and methyl S^2_{calc}

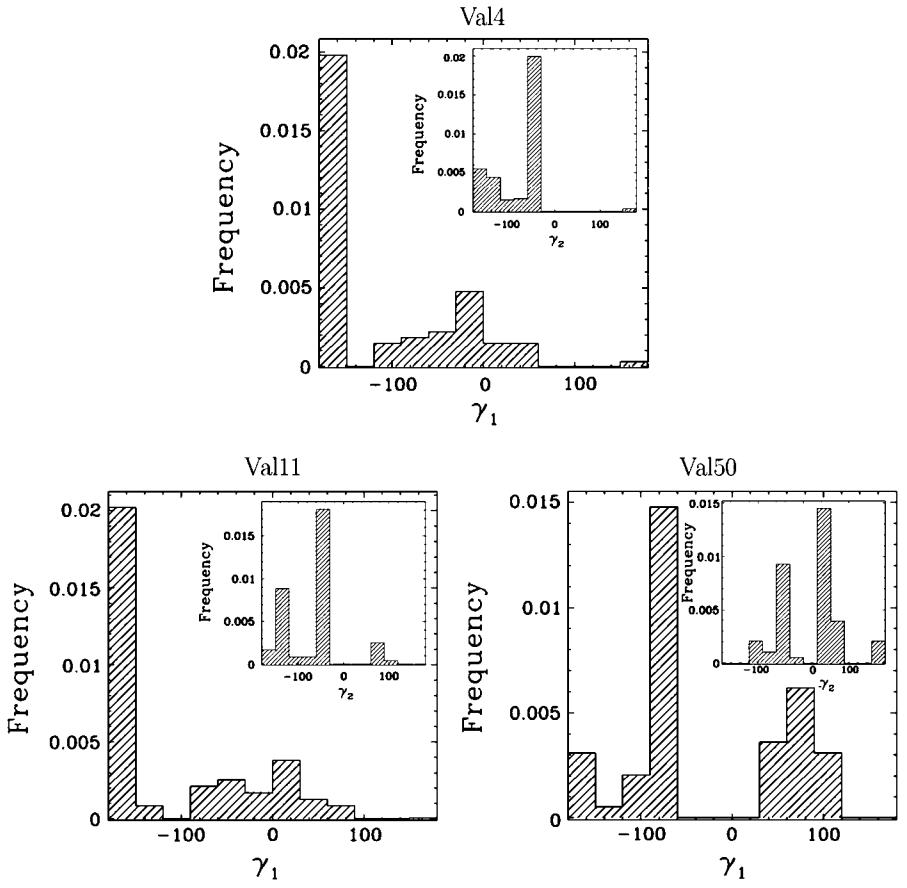

FIGURE 5 Distributions of γ_1 and γ_2 angles for Val^4, Val^{11}, and Val^{50} in FNfn10 reveal that these amino acids visit an average of 4–5 other rotamers. The distributions of γ_2 angles are shown inside the distributions of the γ_1 angles. Averaging over the rotameric states explains these amino acids' unusually low ^3J data, even though only small-scale backbone fluctuations are detected in FNfn10.

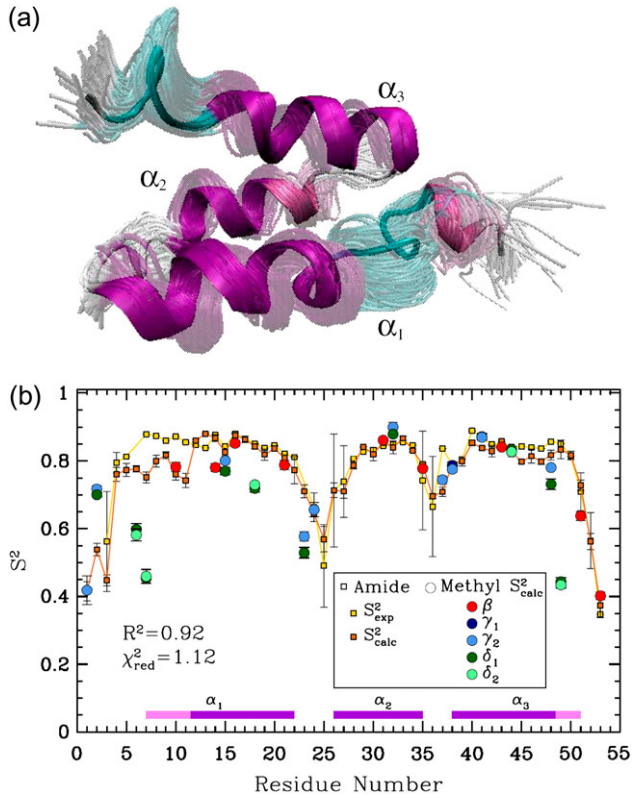

FIGURE 6 (*a*) ALB8-GA Conformations with energy no higher than 5 kcal/mol from the equilibrated solution structure, shown as opaque, are drawn superimposed in transparent representation. (*b*) Calculated amide S^2_{calc} data (*orange squares*), are compared to NMR S^2_{exp} data (*yellow squares*). PEM-obtained methyl S^2_{calc} data are shown in colored circles (no NMR data are available for comparison). Horizontal bars on the *x* axis show the position of the three α-helices on the amino acid sequence of ALB8-GA. The parts of these bars drawn in lighter colors indicate amino acids that are found in unfolded configurations as well.

measured over the obtained ensemble. Amide S^2_{calc} and S^2_{exp} data (42) for ALB8-GA agree with a Pearson correlation of 92% and reduced χ^2 of 1.12. Since NMR methyl S^2 data are currently not available for comparison, in Fig. 6 *b* we show our prediction of methyl S^2 data as obtained by PEM.

The ensemble drawn in Fig. 6 *a* shows that the second α-helix of ALB8-GA, α_2, is tightly packed between the other two helices, α_1 and α_3. Fig. 6 *b* shows that obtained backbone fluctuations of α_2 are small (amide S^2 data >0.8). This result supports the loss of conformational flexibility resulting from selective pressure on α_2, which has evolved to bind human serum albumin with high affinity (42).

In contrast, we observe disorder in the N-terminus of α_1. We find that amino acids Leu^7-Lys^{11} located at the beginning of the α_1 helix of the solution structure of ALB8-GA (31) are highly mobile. These amino acids' high fluctuations can be seen in Fig. 6 *b*. Moreover, we find that Leu^7-Lys^{11} can populate both helical and coil configurations. Indeed, while occasionally populating helical configurations in the PEM-obtained ensemble, these amino acids have a high probability to visit unfolded coil-like configurations.

The low helical content of these amino acids in the PEM-generated ensemble can be seen in Fig. 7 *a*. Fig. 7 *a* shows a square symmetric matrix where a blue square at position (i, j) indicates the presence of a contact between amino acid *i* and amino acid *j*, and a red square indicates the formation of a hydrogen bond between amino acids *i* and *j*. Fig. 7 *a* contrasts the contacts and hydrogen bond network as present in the PEM-generated ensemble, shown top left, with the network present in the representative NMR structure of ALB8-GA, shown bottom right. The bottom right half of the map reveals that in the NMR structure hydrogen bonds are present for amino acids Leu^7-Lys^{11} to be in helical configurations. On the other hand, the top left half of the map shows both the scarcity and the low probabilities for hydrogen bonds in this region, indicating that amino acids Leu^7-Lys^{11} visit coil-like configurations in the PEM-generated ensemble with high probability.

The relative populations of helical and coil configurations visited by amino acids Leu7-Lys^{11} can be quantified by measuring the probabilities of the N-terminus amino acids Leu7-Ala^{21} to be in helical configurations in the ALB8-GA ensemble obtained by PEM. Secondary structure assignment for these amino acids on every conformation of the ensemble is computed with STRIDE (43). The measured probabilities are shown in Table 1(b). We have compared these probabilities with the helicity scores produced by Agadir (44), a program that predicts the helical behavior of polypeptide chains given only amino acid sequence information. The complete amino acid sequence of Leu^7-Ala^{21} is shown in Table 1(a). The helicity scores predicted by Agadir are shown in Table 1(c).

The helicity scores predicted by Agadir agree with our prediction that amino acids Leu^7-Lys^{11} of α_1 have lower probabilities of being found in helical configurations in the native state of ALB8-GA compared to amino acids Lys^{12}-Lys^{19}. This can be seen in Fig. 7 *b*, where we plot and correlate the

TABLE 1

(*a*)	L	K	N	A	K	E	D	A	I	A	E	L	K	K	A
(*b*)	0.01	0.10	0.60	0.85	0.88	0.92	0.90	0.96	0.99	1.00	0.94	0.90	0.80	0.70	0.45
(*c*)	4.7	4.6	3.0	14.4	14.4	15.2	15.8	23.5	24.7	24.9	24.4	22.9	19.9	14.6	11.2

The ALB8-GA sequence of amino acids 7–21 is shown in row *a*. The probability of each amino acid to be part of the first α-helix in the ALB8-GA ensemble obtained by PEM is measured over the ensemble conformations and shown in row *b*. The helicity scores predicted for each amino acid by Agadir (44) are shown in row *c*.

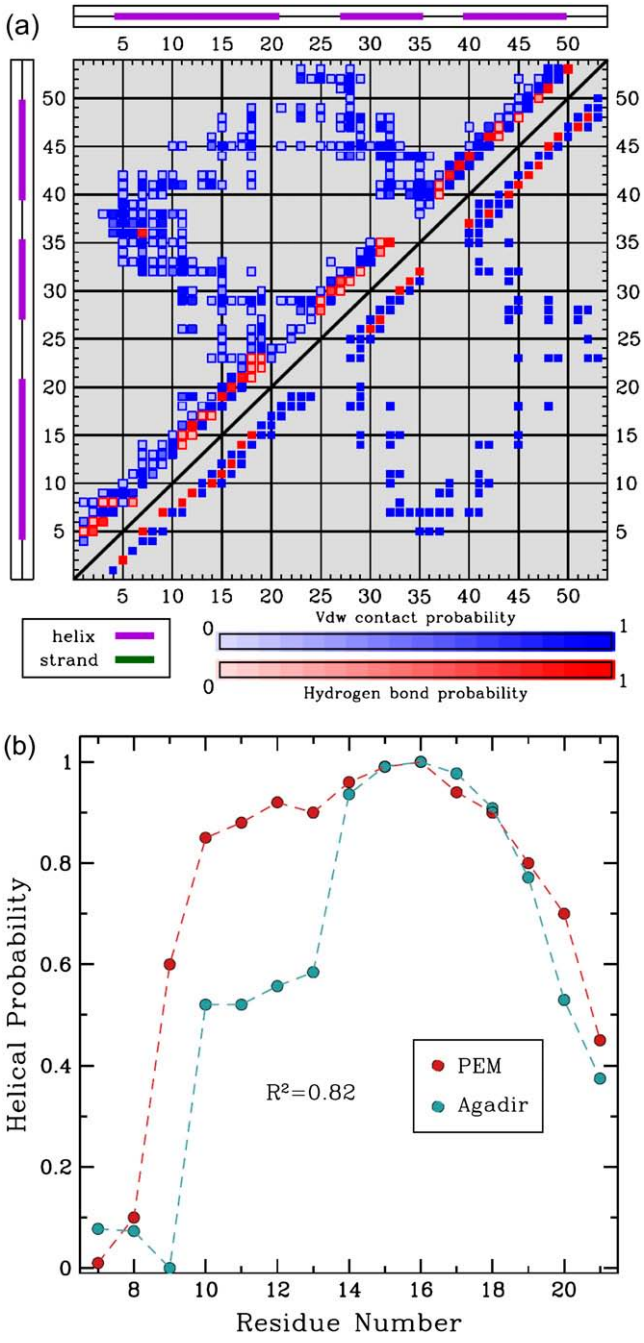

FIGURE 7 (*a*) The contact map is drawn as a 53 × 53 square symmetric matrix (there are 53 aa in ALB8-GA). The formation of a contact between amino acids *i*, *j* is indicated with a blue square at position (*i*, *j*). The formation of a hydrogen bond between *i*, *j* is indicated with a red square at position (*i*, *j*). Shades of blue and red indicate different formation probabilities, with dark blue and dark red indicating a probability of 1, and lighter shades indicating lower probabilities. The top left half of the matrix shows the formation probabilities of contacts and hydrogen bonds in the PEM-generated ensemble. For reference, the bottom right of the matrix shows the contacts and hydrogen bonds in the representative NMR structure of ALB8-GA. The hydrogen bonds in the NMR structure indicate that amino acids Leu^7-Lys^{11} are in helical configurations. The PEM-generated map shows that there are either missing or less probable hydrogen bonds in this region, indicating that Leu^7-Lys^{11} visit unfolded configurations in the PEM-generated ensemble. (*b*) The probabilities for amino acids Leu^7-Ala^{21} to be part of α_1 are shown in red. These probabilities are measured over the ensemble conformations obtained by PEM. The secondary structure assignment for each conformation of the ensemble is computed with the STRIDE program (43) in the Tcl/TK environment of VMD (48). The normalized helicity scores predicted for each amino acid by Agadir (44) are shown in blue.

probabilities measured over the PEM-obtained ensemble with the Agadir-predicted scores. Although the comparison with the Agadir-predicted scores can only be interpreted at a qualitative level (the two data sets measure different quantities), the Pearson correlation with these scores is interestingly high, 82%. This agreement further supports our claim that these five amino acids (Leu^7-Lys^{11}) at the beginning of the α_1 helix in ALB8-GA have indeed a high probability to visit unfolded configurations under native conditions.

Since helix/coil transitions happen on timescales longer than nanoseconds (45), the unfolding observed for amino acids Leu^7-Lys^{11} cannot be detected by the NMR amide S^2_{exp} data (42). The native state ensemble obtained by PEM for ALB8-GA may contain additional information to what is present in the available NMR data. It would be interesting to devise wet-lab experiments that can observe native fluctuations of α_1 over longer timescales. In this particular case, by capturing helix-coil transitions, such experiments could allow to test our prediction of low helical content for Leu^7-Lys^{11}.

DISCUSSION AND CONCLUSION

In summary, we have shown that PEM fully characterizes native local fluctuations of small- to medium-size proteins at atomistic detail. The remarkably good agreements between the available NMR data for these proteins and the thermodynamic properties measured over the PEM-obtained ensembles show that PEM efficiently characterizes native state ensembles in detail, at least for the proteins presented here.

Unlike in trajectory-based simulation techniques, the native conformations obtained by PEM are not correlated to one another. It is this feature that gives PEM its inherent lack of timescale limitations and makes the method intrinsically parallel. The massive parallelism together with the efficient sampling and geometric techniques employed to generate each all-atom conformation of the native state, make PEM an efficient method to obtain extensive native state ensembles of thousands of conformations.

It is worth stressing that the agreement obtained between PEM-calculated and experimental order parameter and scalar coupling data is still a challenge for MD or MC simulation techniques, since slow side-chain rotations may take up to milliseconds (46). In addition, the rotameric averaging measured in the scalar couplings may take from picoseconds to few hundredths of a second (47).

As a sampling-based approach with no inherent timescale limitations, PEM can complement current simulation techniques in highlighting structural and thermodynamic properties

of the native state. In particular, as demonstrated for ALB8-GA, PEM can also complement experimental techniques and formulate hypotheses that can be tested through wet-lab experiments.

It is worth stressing that PEM is primarily intended for application on proteins with nonconcerted motions, as for instance the proteins studied in this article. By obtaining conformations of one segment at a time while maintaining the rest of the protein in a reference conformation, as a first-order approximation method, PEM does consider the possibility of correlated motions of segments far away in sequence. We are currently investigating higher-order approximations (25) to extend PEM to proteins with concerted motions and, more generally, to larger and more complex systems. The results presented in this work lead us to believe that PEM represents a significant first step toward improving our characterization and understanding of protein function at a microscopic scale.

This work is supported by National Science Foundation (C.C., Career grant No. CHE-0349303 and L.E.K. and C.C. grant No. CCF-0523908), the National Institutes of Health (L.E.K. grant No. GM078988), the Welch Foundation (C.C. Norman Hackermann Young Investigator award and grant No. C-1570), and the Sloan Foundation (L.E.K.). A.S. is partly supported by a training fellowship from the Nanobiology Training Program of the W. M. Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (National Institutes of Health grant No.1 R90 DK71504-01). This work was supported in part by the Rice Computational Research Cluster funded by the National Science Foundation under grant No. CNS-0421109 and grant No. CNS-0454333, and a partnership between Rice University, AMD, and Cray.

REFERENCES

1. Kay, L. E. 2005. NMR studies of protein structure and dynamics. *J. Magn. Reson.* 173:193–207.
2. Karplus, M., and J. Kuriyan. 2005. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. USA.* 102:6679–6685.
3. Eisenmesser, E. Z., O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern. 2005. Intrinsic dynamics of an enzyme underlies catalysis. *Nature.* 438:117–121.
4. Peters, G. H., T. M. Frimurer, and O. H. Olsen. 1999. Molecular dynamics simulations of protein-tyrosine phosphatase 1B. I. Ligand-induced changes in the protein motions. *Biophys. J.* 77:505–515.
5. Balabin, I. A., and J. N. Onuchic. 2000. Dynamically controlled protein tunneling paths in photosynthetic reaction centers. *Science.* 290: 114–117.
6. Schnell, J. R., H. J. Dyson, and P. E. Wright. 2004. Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.* 33:119–140.
7. Smith, G. R., M. J. E. Sternberg, and P. A. Bates. 2005. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J. Mol. Biol.* 347:1077–1101.
8. Igumenova, T. I., K. K. Frederick, and A. J. Wand. 2006. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chem. Rev.* 106:1672–1699.
9. Lee, A. L., and A. J. Wand. 2006. Microscopic origins of entropy, heat capacity and the glass transition in proteins. *Nature.* 411:501–504.
10. Lindorff-Larsen, K., R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo. 2005. Simultaneous determination of protein structure and dynamics. *Nature.* 433:128–132.
11. Okamoto, Y. 2004. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graph. Model.* 22:425–439.
12. Singhal, N., C. D. Snow, and V. S. Pande. 2004. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.* 121: 415–425.
13. Jayachandran, G., V. Vishal, and V. S. Pande. 2006. Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. *J. Chem. Phys.* 124: 164902–164914.
14. Chen, J., and C. L. Brooks III. 2005. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. *J. Comput. Chem.* 26:1565–1578.
15. Bahar, I., and A. J. Rader. 2005. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* 15:1–7.
16. Tai, K., T. Shen, U. Boerjesson, M. Philippopulos, and J. A. McCammon. 2001. Analysis of a 10-ns molecular dynamics simulation of mouse acetylcholinesterase. *Biophys. J.* 81:715–724.
17. Jacobs, D. J., A. J. Rader, L. A. Kuhn, and M. F. Thorpe. 2001. Protein flexibility predictions using graph theory. *Proteins Struct. Funct. Genet.* 44:150–165.
18. Rod, T. H., J. L. Radkiewicz, and C. L. Brooks III. 2003. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc. Natl. Acad. Sci. USA.* 100:6980–6985.
19. Daggett, V. 2000. Long timescale simulations. *Curr. Opin. Struct. Biol.* 10:160–164.
20. Price, D. J., and C. L. Brooks III. 2002. Modern protein force fields behave comparably in molecular dynamics simulations. *J. Comput. Chem.* 23:1045–1057.
21. Hansson, T., C. Oostenbrink, and W. F. van Gunsteren. 2002. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* 12:190–196.
22. Wrabl, J. O., S. A. Larson, and V. J. Hilser. 2001. Thermodynamic propensities of amino acids in the native state ensemble: implications for fold recognition. *Protein Sci.* 10:1032–1045.
23. Best, R. B., and M. Vendruscolo. 2004. Determination of ensembles of structures consistent with NMR order parameters. *J. Am. Chem. Chem. Soc.* 126:8090–8091.
24. Chen, J., H. S. Won, W. Im, H. J. Dyson, and C. L. Brooks III. 2005. Generation of native-like protein structures from limited NMR data, modern force fields and advanced conformational sampling. *J. Biomol. NMR.* 31:59–64.
25. Shehu, A., C. Clementi, and L. E. Kavraki. 2006. Computing protein conformations from a single structure: modeling protein flexibility at equilibrium. *Algorithmica*. In press.
26. Shehu, A., C. Clementi, and L. E. Kavraki. 2006. Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. *Proteins Struct. Funct. Bioinf.* 65:164–179.
27. Canutescu, A. A., and R. L. Dunbrack, Jr. 2003. Cyclic Coordinate Descent: a robotics algorithm for protein loop closure. *Protein Sci.* 12:963–972.
28. Hyberts, S. G., M. S. Goldberg, and G. Wagner. 1992. The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with x-ray structures. *Protein Sci.* 1:736–751.
29. Morton, C. J., D. J. R. Pugh, E. L. J. Brown, J. D. Kahmann, D. A. Renzoni, and I. D. Campbell. 1996. Solution structure and peptide binding of the SH3 domain from human Fyn. *Struct. Fold. Des.* 4: 705–714.
30. Main, A. L., T. S. Harvey, M. J. Baron, and I. D. Campbell. 1992. The three-dimensional structure of the tenth type III module of fibronectin: an insight into RGD-mediated interactions. *Cell.* 71:671–678.
31. Johansson, M. U., M. de Château, M. Wikström, S. Forsén, T. Drakenberg, and L. Björck. 1997. Solution structure of the albumin-binding GA module: a versatile bacterial protein domain. *J. Mol. Biol.* 266: 859–865.

32. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliand, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

33. MacKerell, J. A. D., D. Bashford, M. Bellot, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, B. Schlenkrich, J. C. Smith, R. H. Stote, J. Straub, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.

34. Lipari, G., and A. Szabo. 1982. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* 104:4546–4559.

35. Lipari, G., and A. Szabo. 1982. Protein dynamics and NMR relaxation: comparison of simulations with experiment. *Nature.* 300:197–198.

36. Chou, J. J., D. A. Case, and A. Bax. 2003. Insights into the mobility of methyl-bearing side chains in proteins from $^{3}J_{CC}$ and $^{3}J_{CN}$ couplings. *J. Am. Chem. Soc.* 125:8959–8966.

37. Bevington, P. R., and D. K. Robinson. 2002. Data Reduction and Error Analysis for the Physical Sciences, 3rd Ed. D. Bruflodt and S. J. Cotkin, editors. McGraw-Hill, New York, NY.

38. Clarkson, M. W., and A. L. Lee. 2004. Long-range dynamic effects of point mutations propagate through side chains in the serine protease inhibitor eglin c. *Biochemistry.* 43:12448–12458.

39. Mittermaier, A., and L. E. Kay. 2004. The response of internal dynamics to hydrophobic core mutations in the SH3 domain from the Fyn tyrosine kinase. *Protein Sci.* 13:1088–1099.

40. Carr, P. A., H. P. Erickson, and A. G. Palmer III. 1997. Backbone dynamics of homologous fibronectin type III cell adhesion domains from fibronectin and tenascin. *Struct. Fold. Des.* 5: 949–959.

41. Best, R. B., T. J. Rutherford, S. M. V. Freund, and J. Clarke. 2004. Backbone dynamics of homologous fibronectin type III cell adhesion domains from fibronectin and tenascin. *Biochemistry.* 43:1145–1155.

42. Johansson, M. U., H. Nilson, J. Evenäs, S. Forsén, T. Drakenberg, L. Björck, and M. Wikström. 2002. Differences in backbone dynamics of two homologous bacterial albumin-binding modules: implications for binding specificity and bacterial adaptation. *J. Mol. Biol.* 316:1036–1099.

43. Frishman, D., and P. Argos. 1995. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Genet.* 23:566–579.

44. Muñoz, V., and L. Serrano. 1997. Development of the multiple sequence approximation within the Agadir model of α-helix formation. Comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers.* 41:495–509.

45. Doshi, U. R., and V. Muñoz. 2004. The principles of α-helix formation: explaining complex kinetics with nucleation-elongation theory. *J. Phys. Chem. B.* 108:8497–8506.

46. Ming, D., and R. Brueschweiler. 2004. Prediction of methyl side-chain dynamics in proteins. *J. Biomol. NMR.* 29:363–368.

47. Bax, A., G. W. Vuister, S. Grzesiek, F. Delaglio, A. C. Wang, R. Tschudin, and G. Zhu. 1994. Measurement of homo- and heteronuclear J couplings from quantitative J correlation. *Meth. Enzymol.* 239:79–105.

48. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD—Visual Molecular Dynamics. *J. Mol. Graph.* 14:33–38.